多交互 TTS 面向专业录音再现

*Hiroki Kanagawa**, *Kenichi Fujita**, *Aya Watanabe*, *Yusuke Ijima*

NTT, Inc., Japan

hiroki.kanagawa@ntt.com, kenichi.fujita@ntt.com

Abstract

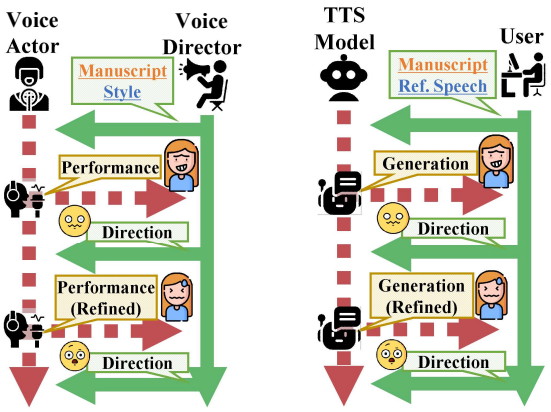
声优导演通常通过提供反馈来迭代精化声优的表现，以达到预期的效果。虽然这种基于反馈的迭代改进过程在实际录音中非常重要，但在文本到语音合成（TTS）中却被忽视了。因此，在初始合成后无法进行细粒度风格调整，尽管生成的语音往往与用户的期望风格不符。为了解决这一问题，我们提出了一种具有多步骤交互的 TTS 方法，允许用户直观且快速地精化生成的语音。我们的方法模拟声优和声优导演之间的关系，建模了 TTS 模型与其用户之间的互动。实验表明，该模型及其相应的数据集能够根据用户的指示进行迭代风格调整，从而展示了其多交互能力。

Index Terms: 文本转语音，富有表现力的语音，说话风格优化

1. 介绍

几乎所有的创造性活动都需要大量的试验和错误才能产生高质量的作品 [1–3]。电影或电视导演通常需要多次拍摄，而摄影师则经常调整角度和相机设置。同样，配音导演通常通过提供反馈来迭代地完善配音演员的表演，以达到期望的说话风格（见图 1(a)）就像演员表演的完善一样 [4,5]。与手工创作过程类似，最近的机器学习任务现在包括用户和系统之间的迭代交互；例如，交互式图像生成 [6–8] 和程序编码 [9,10]。这种趋势正随着大型语言模型 (LLMs) [11,12] 的出现而进一步加速。然而，在语音合成中，这些迭代完善过程尚未得到研究。

最近的文本到语音合成(TTS)技术发展迅速，



(a) 实际记录　　(b) 提出的方法

图 1: 任务概述及我们提出的方法。

实现了自然度与录音相当的人级语音合成 [13,14]。除了自然度外，表现力也有了显著提高，能够再现多种说话风格。对表现力 TTS 的研究将语音分为三个要素：文本、说话者和风格信息。从基于简单的风格 ID 的方法 [15] 以及用原始语音的音高和时长替换的方法 [16, 17] 开始，最近风格信息已使用带有语音嵌入的数据驱动技术进行建模 [18–21]。由于大语言模型（LLMs）的最新进展，可以从说话风格的文字描述中有效地推断出风格信息 [22–24]。然而，这些方法不支持从之前生成的语音反馈，需要 TTS 用户迭代地选择语音提示或文本提示，直到达到所需的合成风格。

为了增强语音合成的适用性，通过实现轻松生成具有所需风格的语音，我们相信引入迭代细化是很有前景的。基于这一想法，我们提出了一种新的由多个方向引导的情感 TTS 方法（见图 1(b)）。为了逐步将 TTS 输出调整为所需的风格，我们将 TTS 模型视为配音演员——类似于配音演员将文本转化为语音——并设计其能够接受额外

*These authors contributed equally.

的文字指导。使用提供的文字指导合成的语音被传递到随后的交互中。通过重复这一交互过程，可以有效地细化合成语音的风格。为了训练我们的模型，我们设计了一个新的数据集来复制实际录音。该数据集包括多个方向文本和配音演员的录音，以模拟风格细化的过程。这些方向文本不仅包含说话风格的操作，还包含语言信息（例如，无声停顿的位置），这与类似的文字提示 TTS 或语音转换方法 [25] 不同。在收集数据时，为了减少对声音导演技能和个性的依赖，文字指导事先准备好了而不是在现场给予。

作为在表达性 TTS 中实现类似人类的交互式迭代风格优化的第一步，我们仅专注于由文本方向引导的语音嵌入操作进行了实验。结果证实了我们的方法能够在不降低自然度的情况下大致将用户的意图方向纳入合成语音中。此外，我们的案例分析还强调了一些遗留问题，如 1) 句子中的具体位置和 2) 处理模糊表达，为进一步改进风格优化提供了重要见解 [1]。

2. 提出的方法

为了模拟语音导演与配音演员之间的互动，我们构建了一个模仿他们实际录音的数据集。在录制该数据集的过程中，导演反复给出表演指示，然后配音演员反映这些指示。为了模拟这一指导循环，我们构建了一个能够生成反映出所给指示的精致语音的模型。我们希望强调的是，风格精修未来将包括语言方面的精修。因此，我们的任务完全不同于使用文本提示 [25] 的语音转换，在后者中只改变说话人的特征而保持语言信息不变。本节解释数据集 (图 2)、骨干 TTS 模型和风格精修器 (图 3)。

2.1. 数据集

为了建模声音导演和配音演员之间的互动，我们构建了一个数据集，其中包含经配音演员反复改进的演讲内容以及声音导演提供的指导（图 2）。

[1] https://ntt-hilab-gensp.github.io/ssw13multiinteractiontts/

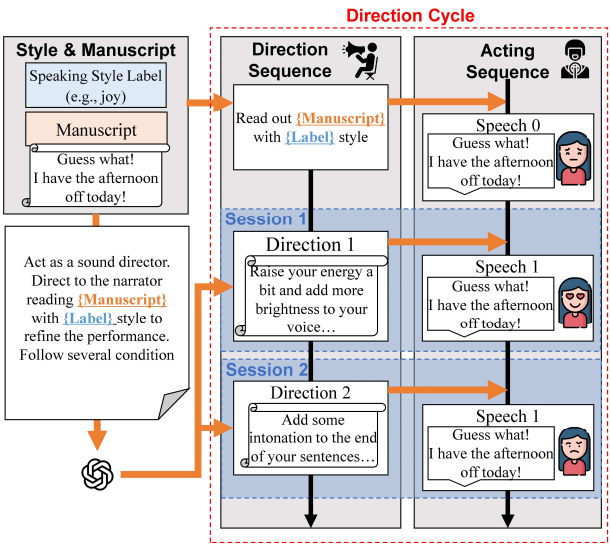


图 2: 数据集收集中的方向循环概述。

表 1: 样式标签、手稿和方向从日语翻译的示例。

	Example
Style Label	Resignedly
Manuscript	I don't have time, okay? It's not like I wanted this either...
Direction 1	Say it in a more restrained tone, with a sigh mixed in. Keep the delivery flat and steady, with little variation in strength.
Direction 2	Let's try speaking so that your intonation gently falls at the end without sharp breaks, letting your words flow smoothly. Keep a slightly laid-back tone in mind.
Direction 3	Place just a slight pause between words, and speak in a tone that conveys resignation rather than urgency; overall, keep your pitch low.

在实际的录音过程中，声导演员以各种方式向配音演员提供表演指导。此类指导可以归类如下：

1. 修改表演言语中的副语言和非语言信息的指示（例如，“说话更开朗一些”，“更多地考虑对方的感受”）。
2. 修改语言信息的说明（例如，“在此词后插入一个无声停顿”，“改变重音位置（或对于声调语言，如日语中的声调重音）”）。
3. 演示性指示，其中导演提供一个表演示例（例如，“跟我做”）。
4. 涉及手势和肢体语言的指示。

虽然构建包含所有这些类别的数据集是最理想的，但创建如此全面的数据集极其复杂，并且存在重大挑战。因此，在本研究中，我们仅构建了涵盖类别 1 和 2 的数据集，这些类别可以单纯以文本形式表达。

此外，在大多数实际的录音过程中，声音导演会反复审查结果并提供指导，以引导配音演员达到预期的效果。然而，复制这一过程在数据集中是困难的，因为指令词汇的多样性和一致性很大程度上取决于导演的专业水平。因此，这种方法不适合用于构建大规模的数据集。

为了可靠地收集大量数据而不依赖声导的技能，我们使用了一个大型语言模型（LLM）生成指示，而没有参考实际录音。这减少了声导个人风格的影响，并且更容易获得一个大规模的数据集同时保持指示的质量。

为了构建包含各种说话风格的数据集，我们准备了与风格类别（例如，困倦的、面向儿童的和紧急的）相匹配的手稿，遵循“TTS 说话风格分类指南”[26]。基于这些说话风格，我们通过为大型语言模型输入文本提示来生成每份手稿的方向（示例可以在我们的演示页面上找到）。该大型语言模型被指示充当声音导演，并生成两到三次迭代方向。为了验证质量，三位日语母语者手动验证了格式和整体质量。与诸如 PromptTTS [22] 等基于文本提示的 TTS 方法不同，我们生成的方向包括更复杂和抽象的文本描述；以往研究中的文本提示仅限于简单的指示，例如高/低音调、快/慢语速。此外，这些方向还涵盖了改变语言信息的指令，例如插入停顿。

2.2. 骨干 TTS 模型

TTS 模型依赖于从目标语音中提取的嵌入。为了获得该嵌入，我们使用了语音编码器和风格标记层（STL）[18]。具体来说，语音编码器利用自监督学习（SSL）语音模型 [27] 来抽取特征表示，然后通过加权求和 [28]、双向 LSTM 和注意力机制 [29] 将这些特征聚合为一个向量，记作 $\boldsymbol{x}$。随后，应用了 STL 后处理 $\boldsymbol{x}$ 以增强稳定性，从而得到了最终的语音嵌入。通过联合训练 TTS 模型与这些模块，所生成的语音嵌入非常适合 TTS 系统。

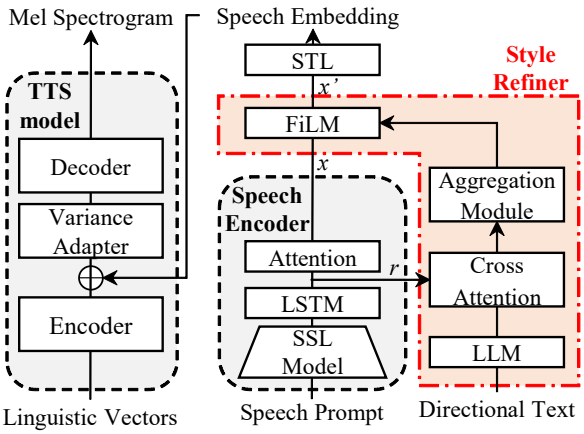


图 3: 提出的模型概述。

2.3. 风格优化器

风格细化器将预细化嵌入 $\boldsymbol{x}$ 细化为细化嵌入 $\boldsymbol{x'}$，使用了方向文本中的信息。风格细化器由三个组件组成：交叉注意力、聚合模块和 FiLM [30]。交叉注意力整合了语音和文本中的信息，如同其他关于文本-语音整合的研究 [31, 32]。为了处理更丰富的语音信息，交叉注意力采用来自语音编码器中间 LSTM 层的输出 $\boldsymbol{r} \in \mathbb{R}^{D_r \times L_r}$ 而不是 $\boldsymbol{x} \in \mathbb{R}^{D_x \times 1}$ 作为输入。这里，D_r、L_r 和 D_x 分别表示 $\boldsymbol{r}$ 的维度和帧数以及 $\boldsymbol{x}$ 的维度。集成后，聚合模块使用与语音编码器相同的注意力机制将表示整合为一个固定维度的嵌入。最后，$\boldsymbol{x}$ 通过一个受聚合表示条件控制的 FiLM 层转换为 $\boldsymbol{x'}$。

风格精炼器是单独训练的，首先训练骨干 TTS 模型。我们从预精炼语音中提取了 $\boldsymbol{x}$ 和 $\boldsymbol{r}$，并从精炼语音中提取了 $\boldsymbol{x'}$。这些提取的表示和相应的方向文本配对并用于训练风格细化器。目标与预测的 $\boldsymbol{x'}$ 之间的 L1 损失被利用。在推理时，风格精进器生成 $\boldsymbol{x'}$，然后传递给 STL 并转换为语音嵌入。此嵌入随后输入 TTS 模型以生成预期根据给定方向精炼的预精炼语音的修改版本。

3. 实验设置

3.1. 数据集

我们使用了三个日语 22 kHz 数据集：交互式、非交互式和大型内部数据集。交互式和非交互式数据集来自同两名配音演员（一名女性和一名男性）。该交互式数据集特别为本研究设计，如第 2.1 节所述。它包括 13.7 小时的日语语音，由从 1,606 个方向循环中收集的 5,639 条陈述组成。从此数据集中，选择了 104 个方向循环（所有 52 种风格 ×2 名配音演员）用于验证，并另选了 104 个进行测试。使用 ChatGPT-4o [33] 准备了指导文本，风格标签是从 52 种说话风格中选出的（排除了原本指南 [26] 中 55 种说话风格中的中性、尖叫和大笑）。在录制过程中，为了真实模仿，未来的方向被隐藏起来不告诉配音演员，并且按需依次向他们提供必要的指导。

非交互式数据集是不包含交互会话的标准 TTS 数据集，包括 43.6 小时的语音数据，共含 42,039 个语句。该数据集包含了遵循指南 [26] 的 55 种演讲风格的表演语音。大型内部数据集包括 8K 小时的日语语音。

为了提高风格细化器对方向多样性的鲁棒性，我们采用了四种基于大语言模型的增强方法。第一种是*简单*；它通过大语言模型总结原始方向，假设用户无法提供详尽的方向。另外三种则是通过 Maini 等人提出的提示来实现的。即使用他们的提示模板：*简单*、*硬*和*中等风格*（不包括*问答风格*）[34]。由于数据集是日语，我们使用了这些模板的翻译版本。我们使用了 Llama-3.1-Swallow-8B-Instruct v0.2[2] 作为大语言模型，增强的方向最终每个原始方向产生了七个结果（两个带有*简单*，*硬*和*介质*，一个带有*简单*提示）。

3.2. 训练条件

TTS 模型基于 FastSpeech2 [35]，遵循之前工作中描述的实现方法 [36]。TTS 模型的输入是包含音素和重音信息的 303 维语言向量。输出是一个带有 10 ms 帧移的 80 维对数梅尔频谱图。我们整合了一个在 ReazonSpeech[3] 上预训练的 HuBERT 基础模型，并冻结了其参数。该模型将每个 16 kHz 原始音频样本转换为一个 768 维向量序列，然后将其变换为固定长度的 384 维*语音嵌入*。语音波形由 HiFi-GAN (V1) [37] 生成。作为风格优化器，我们使用了在日语数据集上微调的 Gemma2 大型语言模型 [4]。不同的大型语言模型分别用于风格优化（Gemma2）、方向文本生成（ChatGPT-4o）和增强（Llama-3.1-Swallow-8B-Instruct v0.2），以防止由于从单一模型的有偏知识训练而引起的性能崩溃，如在 [38] 中所报告的。

所提出的模型按照以下四个步骤进行训练。

1. 使用大型内部数据集训练主干 TTS 模型历时 60 万步。
2. 骨干 TTS 模型以与 [39, 40] 相同的方式使用 GAN 进行进一步训练，额外训练了 200K 步，固定学习率为 1×10^{-4}，以提高合成语音的自然度。
3. 模型随后在 30K 步中使用交互和非交互数据集进行了微调，以改进每位配音演员的说话风格再现。损失函数和学习率与上一步相同。
4. 样式精炼器通过交互数据集进行训练。训练步数为 10K 步，使用 AdamW 优化器 [41]，带有 4K 预热步数。

4. 主观评估

我们进行了三项主观评估以确认所提出方法的有效性。这些评估检查了迭代风格细化、细化精度和自然度。对于此次评估，我们将测试数据集（52 种说话风格）分为五个风格组，具体如 [42] 所示。这五组可以解释为恐惧、快乐、愤怒、悲伤和惊讶的群体。

4.1. 演讲准备

语音样本是在四种条件下生成的：*相同*，*演员引导的 (预言机)*，*单次射击 (我们的)* 和*迭代 (我们*

[2] https://huggingface.co/tokyotech-llm/Llama-3.1-Swallow-8B-Instruct-v0.2

[3] https://huggingface.co/rinna/japanese-hubert-base

[4] https://huggingface.co/rinna/gemma-2-baku-2b-it

表 2: 语音生成条件在一个方向周期跨越三个会话中的表现，说明了不同的方法（相同、演员引导、单次和迭代）如何利用语音提示和方向。

Conditions	第 1 节			第 2 节			第 3 节		
	输入		Output	输入		Output	输入		Output
	Speech Prompt	Direction 1	Speech 1	Speech Prompt	Direction 2	Speech 2	Speech Prompt	Direction 3	Speech 3
Identical	Recorded-0	-	**相同-0**	Recorded-0	-	**相同-0**	Recorded-0	-	**相同-0**
Actor-Guided (oracle)	Recorded-1	-	Guided-1	Recorded-2	-	Guided-2	Recorded-3	-	Guided-3
Single-shot (ours)	Recorded-0	✓	Single-1	Recorded-1	✓	Single-2	Recorded-2	✓	Single-3
Iterative (ours)	Recorded-0	✓	迭代-*1*	迭代-*1*	✓	迭代-*2*	迭代-*2*	✓	迭代-*3*

的*)*，如表 2 所示。在语音生成过程中，我们使用了两种类型的语音提示：实际录制的语音（记录的-*N*）和生成的语音（迭代-*N*，相同-*0* 和引导式-*N*），其中 *N* 表示该语音取自方向循环的第 *N* 轮。

在迭代（我们的）和单次射击 *(*我们的*)* 条件下，我们的风格细化器接收了语音提示（之前会话中迭代或非迭代生成的）以及相应的方向文本。在相同和 *Actor*-引导的（预言机）条件下，没有使用我们的风格细化器来生成语音（即正常 TTS 生成），而是使用实际录制的语音提示。

4.2. 迭代风格细化

我们进行了一次迭代风格精炼的主观评估。参与者看到了在三种条件下最多三次精炼迭代的伪方向循环：表 2 中的相同、演员引导（我们的）和迭代（我们的）。每个伪方向循环对应于测试数据中的一个周期（图 2）。参与者看到了方向文本，然后是参考演讲和测试演讲，形成一个伪方向循环（详情见表 3）。随后，参与者用五点 MOS 测试为每一对打分，使用的方向文本如下：1 - 没有变化；2 - 小幅修改；3 - 只有整体一致；4 - 反映导演意图的调整；5 - 充分体现导演意图的重大改变。在评估之前，参与者被提供了与每个分数相对应的具体示例。例如，3 分（“只有整体一致”）解释为演讲总体上符合所要求的情感基调或风格，但缺乏具体的改进之处的情况。相同条件通过评定相同的语音对来验证我们的评价基准。我们期望方向能在实际录音中得到正确反映，因此演员引导（预言机）条件代表了我们方法的上限。

评估通过 82 名参与者在一个众包界面中进行。对于评估，我们为每种方法选择了 16 个方向

表 3: 来自方向循环中 第 *4.2* 节和 *4.3* 节评估的成对语音。

方法	Identical	Actor-guided	Single-shot	Iterative
风格优化器	-	-	✓	✓
Session 1 Ref. Speech	**相同-0**	**相同-0**	**相同-0**	**相同-0**
测试演讲	**相同-0**	Guided-1	Single-1	迭代-*1*
Session 2 Ref. Speech	**相同-0**	Guided-1	Guided-1	迭代-*1*
测试演讲	**相同-0**	Guided-2	Single-2	迭代-*2*
Session 3 Ref. Speech	**相同-0**	Guided-2	Guided-2	迭代-*2*
测试演讲	**同型-0**	Guided-3	Single-3	迭代-*3*

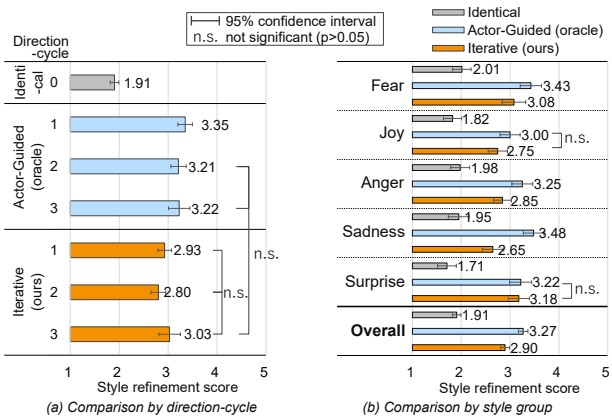


图 4: 主观评估结果对于迭代风格细化。(*a*) 比较了每种方法关于方向-周期。(*b*) 将它们分解为五个风格组并显示总体结果。

循环，即每个测试数据集中的语音演员的每个子组选取三个或四个循环。结果如图 4 (a) 所示。迭代（我们的）由于我们的风格优化器的有效性显著优于相同。另一方面，迭代 *(*我们的*)* 稍微逊色于演员引导的（预言机）。这些结果表明我们提出的方法实现了与预期改进相符合的风格细化。演员引导的（预言机）和迭代 *(*我们的*)* 之间的差距表明与期望的风格仍存在差异。迭代 *(*我们的*)* 条件也显示在方向周期中没有显著分数变化，这意味着无论之前的改进次数多少，方向都会反映在合

成语音中。

图 4(b) 将汇总结果分解为风格组，并显示了总体结果。迭代（我们的）在所有五个风格组中的得分都低于 *Actor*-引导的（预言机）。虽然在“快乐”和“惊讶”风格组中两种方法之间没有发现显著差异，但在“总体”以及其他风格组中它们之间的差异是显著的。这些结果表明，如图 4(a) 之前所示，迭代 *(我们的)* 在风格细化方面仍有改进的空间。

4.3. 风格精修精度

然后我们评估了改进是否真正符合给定的方向。参与者被呈现文本方向，然后连续展示了预改进和改进后的演讲。随后，参与者通过一个从 −3 (改进偏离了方向) 到 3 (改进符合方向) 的七点 MOS 测试来评估演讲与给定文本的一致性差异。演讲是在表 2 中的单次射击 *(我们的)* 条件下生成的，以及对应于表 3 中单次射击 *(我们的)* 条件下的参考和测试演讲的预优化和优化演讲。

我们准备了两种条件：匹配和随机。在前一种条件下，向参与者展示的方向实际上是用于细化的，在后一种条件下，展示的方向是从测试数据中随机选取的。每位配音演员选择了 50 个会话（五个子组中的每个子组有 10 个会话）。每个会话被评估了三次：一次是在匹配条件下的评估，两次是在随机条件下的评估。为了涵盖句子多样性，随机条件使用两个不同的方向进行评估。参与者人数为 110 人，通过相同的众包界面收集。此外，为了详细分析，我们使用 ChatGPT o3-mini 将随机条件分类为随机 *(相似)* 和随机 *(不同)*。我们要求 ChatGPT o3-mini 对原始用于优化的方向与随机选择的一个方向之间的相似性进行评分，评分为从 1（不相似）到 5（非常相似）。根据这些预测分数，我们将得分在 2 分或以下的句子视为随机 *(不同)*，将得分在 3 分或以上的句子视为随机 *(类似)*。

图 5 展示了使用参与者获得的原始分数，每种风格和每种条件下风格精炼准确率的小提琴图。匹配到在所有风格组中显著优于{随机 *(不同类)*。随机 *(相似)* 与匹配的性能相当。这些结果表明，即使用于风格精炼的方向与给听众的方向不完全匹配，但如果它们的方向相似，则可以获得相当的性能。另一方面，当给听众的方向句子与用于风格细化的完全不同（即，随机 *(不同)*）时，分数明显比匹配的和随机 *(相似)* 要差，且分数分布偏向较低。这些结果显示我们方法实现的细化与给定方向一致。

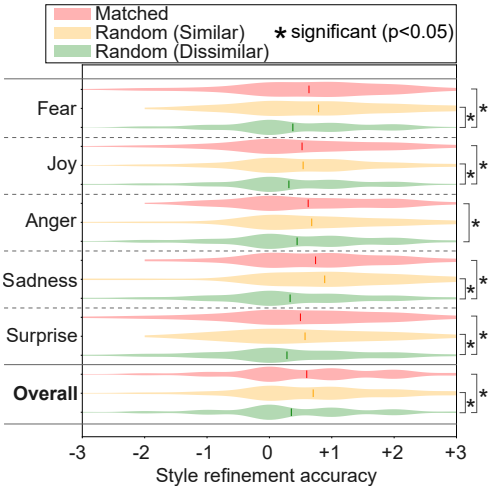


图 5: 主观评估结果关于风格精修准确性。“总体”结果是从所有评估分数中获得的，未进行风格组分类。

4.4. 自然性

我们最终使用五点量表（1：非常不自然，5：非常自然）的 MOS 分析了迭代生成对自然度的影响。我们比较了三种条件：单次射击 *(我们的)*、迭代 *(我们的)* 和演员引导（预言机），如表 2 所示。参与者人数为 168 人，通过相同的众包界面招募。我们为每种条件选择了 15 个句子，即每个测试数据子组中每位配音演员的三个句子。

结果如图 6 所示。所有条件，即演员引导的（预言机），单次射击 *(我们的)* 和迭代 *(我们的)*，在自然度方面均无显著差异。这意味着我们方法中明确的迭代细化不会降低自然度。

5. 讨论

如 Sect. 2.1 中所述，我们的数据集包含先前基于文本提示的 TTS 方法中未包含的困难方向。我们认为这解释了 Sect. 4.2 中的总体分数相对较低的原因；即使在非迭代设置下，它仍然保持在三

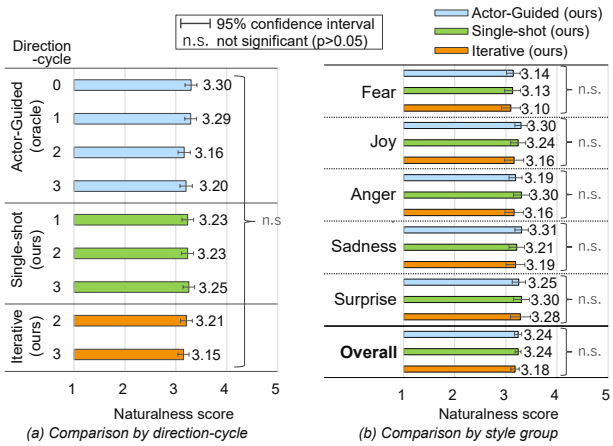


图 6: 主观评估自然度的结果。类似于图 4，*(a)* 对所有方法在方向-周期上的比较。*(b)* 展示了从所有评估数据中获得的风格组和“总体”结果，不包括风格组分类。

左右 (*Actor-*引导的 *(*预言机*)*)。

表 4 列出了实验中得分最高和最低的三个示例，这些示例来自测试数据（参见第 4.2 节）。本节通过检查数据集中的具体示例来讨论问题并为未来的工作提出解决方案。

5.1. 风格建模

本研究主要集中在通过迭代环节进行风格精炼。然而，我们发现一些示例，特别是那些得分较低的示例，包含难以用单一基于语音嵌入的 TTS 系统实现的指令。具体来说，有些示例包含了指定特定单词或手稿中位置（例如，“在开头”，“在结尾”和“对于‘如何做‘部分”）的指令。由于全局嵌入通常不能改变句子内特定单词的表现形式，因此我们的方法很难通过精炼来反映这些指令。一个有希望的解决方案是使用细粒度的方法，如 NaturalSpeech 3 [43] 或 VALL-E [44] 作为 TTS 模型。

5.2. 语言内容建模

在给出行动指令时，支持改变语言信息也很重要。事实上，一些指导包括了对这些特征的说明（例如，“屏住呼吸”和“在词语之间只做一个短暂的停顿”）。然而，我们当前的模型仅优化语音风格同时保持语言信息不变。为了实现这种支持，

表 4: 交互式风格细化评估中，排名前三和最末三样本的方向为交互式 *(*我们的*)*。下划线的 和**粗体**部分分别代表 特定单词或位置的说明 和**更改语言信息的说明**。分数来自第 *4.2* 节。

Score	Direction (Translated from Japanese)
4.67	Try speaking bit faster overall. Especially in 后半部分, picking up the pace slightly will help emphasize the joy of discovery
4.22	在行的开头, try **喘了口气**, as if you're slightly out of breath. This will help convey a sense of excitement.
4.20	Slow down the speaking pace a little, and try to let the sound fade out 在结尾处, as if it's gently disappearing.
2.30	**屏住呼吸** slightly 在结尾处, as if swallowing your words. Make it sound more desperate and urgent.
2.30	Lower your tone 在结尾处 and let your voice fade out weakly. This will create a worn-out impression.
2.08	Extend the ending slightly to add a gentle, inviting nuance. 对于“这样做怎么样”的部分, speak softly as if warmly encouraging the listener.

不仅需要细化副语言信息，还需要从指令中细化语言信息。一个有希望的方法是将 LLM 用于语言信息修改纳入我们的提议方法中。

5.3. 主观评价的设计

本文中的评估考察总体印象。鉴于这是对我们提案的初步评估，这一点是合理的。然而，由于这种评估方法可能无法完全捕捉到细微的风格差异，因此第 4.2 节中演员引导（预言机）条件下的相对较低分数是可以理解的。最近的文字转图像/视频生成任务也面临着相同的评估难题。正在研究一些能够捕获详细生产差异以评估与文本提示 [45, 46] 一致性的方法。同样，需要进一步探索更加细致和合适的评估方法来推进我们的工作。

6. 结论

本文提出了一种 TTS 系统的方法，可以模拟声优导演与配音演员之间的互动。主观评估表明，我们提出的方法实现了某种程度上符合用户指导的迭代风格优化。此外，对几个实验的分析为进一步提升我们的方法以实现人类级别的风格优化提供了建议。未来的工作将涉及进一步细化风格，不仅通过语音嵌入的操作，还包括融合语言信息（例如，静默停顿的位置、重音或音调强调）。对于这项任务的有效评估方法的研究仍然是一个持续的挑战。作为构建风格优化框架的一个初始步骤，我们研究了一种依赖说话人的风格优化器，并计划将这种方法扩展到独立于说话人的情景。

7. References

[1] K.-C. Tönnsen, "The relevance of trial-and-error: Can trial-and-error be a sufficient learning method in technical problem-solving-contexts?" *Techne Series - Research in Sloyd Education and Craft Science A*, vol. 28, no. 2, p. 303 – 312, Apr. 2021.

[2] C. D. Güss, S. Ahmed, and D. Dörner, "From da Vinci' s flying machines to a theory of the creative process," *Perspectives on Psychological Science*, vol. 16, no. 6, pp. 1184–1197, 2021.

[3] M. Botella, F. Zenasni, and T. Lubart, "What are the stages of the creative process? What visual art students are saying." *Frontiers in Psychology*, vol. 9, 2018.

[4] N. J. Glikpoe and I. Horsu, "From page to stage: The director' s interpretation and picturization of a script," *Advanced Journal of Theatre and Film Studies*, vol. 1, no. 1, pp. 36–42, 2023.

[5] A. Schmidt and A. Deppermann, "Showing and telling—How directors combine embodied demonstrations and verbal descriptions to instruct in theater rehearsals," *Frontiers in Communication*, vol. 7, p. 955583, 2023.

[6] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," arXiv, 2022, arXiv:2204.06125.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *Proc. CVPR*, pp. 10 684–10 695, 2022.

[8] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," *Proc. NeurIPS*, vol. 35, pp. 36 479–36 494, 2022.

[9] J. Austin, A. Odena, M. I. Nye, M. Bosma, H. Michalewski, D. Dohan, E. Jiang, C. J. Cai, M. Terry, Q. V. Le, and C. Sutton, "Program synthesis with large language models," arXiv, 2021, arXiv:2108.07732.

[10] E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, Y. Zhou, S. Savarese, and C. Xiong, "CodeGen: An open large language model for code with multi-turn program synthesis," *Proc. ICLR*, pp. 284–295, 2023.

[11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin *et al.*, "Training language models to follow instructions with human feedback," *Proc. NeurIPS*, vol. 35, pp. 27 730–27 744, 2022.

[12] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts *et al.*, "PaLM: Scaling language modeling with pathways," *JMLR*, vol. 24, no. 240, 2023.

[13] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," arXiv, 2021, arXiv:2106.15561.

[14] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, 2023.

[15] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "Model architectures to extrapolate emotional expressions in DNN-based text-to-speech," *Speech Communication*, vol. 126, pp. 35–43, 2021.

[16] M. Aylett, D. Braude, C. Pidcock, and B. Potard, "Voice puppetry: Exploring dramatic performance to develop speech synthesis," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 117–120.

[17] E. Van De Vreken, K. Richmond, and C. Lai, "Voice puppetry with FastPitch," in *Proc. Interspeech*, 2022, pp. 5219–5220.

[18] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style Tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *Proc. ICML*, vol. 80, pp. 5180–5189, 2018.

[19] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," *Proc. ICASSP*, pp. 5911–5915, 2019.

[20] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," *Proc. Interspeech*, pp. 4440–4444, 2019.

[21] J. Zadi, H. Seut, B. van Niekerk, and M.-A. Carbonneau, "Daft-Exprt: Cross-speaker prosody transfer on any text for expressive speech synthesis," *Proc. Interspeech*, pp. 4591–4595, 2022.

[22] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "PromptTTS: Controllable text-to-speech with text descriptions," *Proc. ICASSP*, 2023.

[23] D. Lyth and S. King, "Natural language guidance of high-fidelity text-to-speech with synthetic annotations," arXiv, 2024, arXiv:2402.01912.

[24] P. Peng, P.-Y. Huang, S.-W. Li, A. Mohamed, and D. Harwath, "VoiceCraft: Zero-shot speech editing and text-to-speech in the wild," *Proc. ACL*, pp. 12 442–12 462, 2024.

[25] Z.-Y. Sheng, L.-J. Liu, Y. Ai, J. Pan, and Z.-H. Ling, "Voice attribute editing with text prompt," *TASLP*, pp. 1–12, 2025.

[26] Japan Electronics and Information Technology Industries Association, "The guidelines for TTS speaking style classification (IT-4012)," 2021, (In Japanese). [Online]. Available: https://www.jeita-speech.org/standard/standard_4012.html

[27] K. Fujita, T. Ashihara, H. Kanagawa, T. Moriya, and Y. Ijima, "Zero-shot text-to-speech synthesis conditioned using self-supervised speech representation model," in *Proc. ICASSP Workshops (ICASSPW)*, 2023.

[28] S. Chen, Y. Wu, C. Wang, S. Liu, Z. Chen, P. Wang *et al.*, "Why does self-supervised learning for speech recognition benefit speaker recognition?" in *Proc. Interspeech*, 2022, pp. 3699–3703.

[29] C. Raffel and D. P. W. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," in *Proc. ICLR Workshop*, 2016.

[30] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, vol. 32, no. 1, 2018.

[31] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Proc. Interspeech*, 2019, pp. 3569–3573.

[32] L. Sun, B. Liu, J. Tao, and Z. Lian, "Multimodal cross- and self-attention network for speech emotion recognition," in *Proc. ICASSP*, 2021, pp. 4275–4279.

[33] OpenAI, "Hello GPT-4o," 2024, accessed:2024-02-14.

[34] P. Maini, S. Seto, H. Bai, D. Grangier, Y. Zhang, and N. Jaitly, "Rephrasing the web: A recipe for compute and data-efficient language modeling," in *Proc. ICLR Workshop*, 2024.

[35] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, 2020.

[36] C.-M. Chien, J.-H. Lin, C.-y. Huang, P.-c. Hsu, and H.-y. Lee, "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," in *Proc. ICASSP*, 2021, pp. 8588–8592.

[37] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, vol. 33, 2020, pp. 17 022–17 033.

[38] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, "The curse of recursion: Training on generated data makes models forget," arXiv, 2024, arXiv:2305.17493.

[39] H. Kanagawa and Y. Ijima, "Multi-speaker modeling for DNN-based speech synthesis incorporating generative adversarial networks," in *Proc. SSW*, 2019, pp. 40–44.

[40] X. B. Peng, A. Kanazawa, S. Toyer, P. Abbeel, and S. Levine, "Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow," in *Proc. ICLR*, 2019.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[42] Y. Homma, H. Kanagawa, N. Kobayashi, Y. Ijima, and K. Saito, "Expressive text-to-speech synthesis using text chat dataset with speaking style information," *Transactions of JSAI*, vol. 38, no. 3, pp. F–MA7, May 2023, (In Japanese).

[43] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin, D. Yang, Y. Liu, Y. Leng, K. Song, S. Tang, Z. Wu, T. Qin, X.-Y. Li, W. Ye, S. Zhang, J. Bian, L. He, J. Li, and S. Zhao, "NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *Proc. ICML*, 2024.

[44] S. Chen, C. Wang, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *TASLP*, vol. 33, pp. 705–718, 2025.

[45] Y. Liu, L. Li, S. Ren, R. Gao, S. Li, S. Chen, X. Sun, and L. Hou, "FETV: A benchmark for fine-grained evaluation of open-domain text-to-video generation," in *Proc. NeurIPS*, vol. 36, 2023, pp. 62 352–62 387.

[46] M. Otani, R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh, "Toward verifiable and reproducible human evaluation for text-to-image generation," in *Proc. CVPR*, June 2023, pp. 14 277–14 286.